Toxicity and Digital Incivility during Political Deliberation

The Case of the Constitutional Convention in Chile

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalezbustamante@fgga.leidenuniv.nl

Sebastián Rivera
Universidad Mayor
✉ sebastian.riveraa@umayor.cl

Presentation delivered at the Digital Democracy Workshop
Digital Society Initiative, University of Zurich, Zurich, October 24-25, 2024

Table of Contents

Introduction

Introduction

Incivility has become an unintended consequence of social media dynamics. Anonymity and lack of face-to-face accountability often encourage users to engage in more hostile or confrontational behaviour (Theocharis et al., 2020). Our main question is: **To what extent are politicians subject to uncivil comments on social media platforms?**

We leverage data from Chile, specifically from Twitter interactions between users and members of the **Constitutional Convention**, a political body whose primary task was to draft a new Constitution.

- Perspective API for toxicity classification
- GPTs, open-source LLMs and human annotation
- NBRMs to assess the likelihood of receiving toxic messages and insults

Empirical Expectations

Gender Hypothesis

Indeed, there is some evidence that women are more likely to receive toxicity than men (Gorrell et al., 2020; Kim, 2023). Although other research focuses on the attacks faced by women politicians on social media, these studies examine only female accounts, making inferences about gender differences impossible (Erikson et al., 2023; Esposito and Breeze, 2022; Pérez-Arredondo and Graells-Garrido, 2021; Weaving et al., 2023).

Based on this discussion, we propose the following hypothesis:

- **Gender Hypothesis**. Female representatives are significantly more likely than male representatives to receive toxic comments and insults.

Indigenous Hypothesis

Racial and ethnic minorities in particular are subject to comments that exploit long-standing stereotypes and cultural biases (Gorrell et al., 2019) and often interact with other prejudices about the participation of groups or identities traditionally excluded from the exercise of political power. Indeed, Orchard et al. (2023) analysed how indigenous women experienced more hostility than non-indigenous women by the end of the Constitutional Convention in Chile.

Based on this research, we propose the following hypothesis:

- **Indigenous Hypothesis**. Indigenous representatives are significantly more likely than non-indigenous to receive toxic comments and insults.

Ideology Hypothesis

Politicians at the extremes of the ideological scale tend to evoke stronger emotional reactions, which can amplify the volume of toxic comments directed at them. Indeed, previous research has suggested that politicians associated with extreme ideological positions are more frequently attacked and receive higher toxicity levels on social media (Theocharis et al., 2020).

Based on this discussion, we propose the following third hypothesis:

- **Ideology Hypothesis**. Being ideologically extreme increases the probability of receiving toxic comments and insults.

Data and Methods

Data and Toxicity Classification

Novel dataset

Jan 1, 2021 – Dec 31, 2022

~31.8M messages (>132K JSON files, ~56GB)

~9.2M interactions with Convention members ✔

We applied **Perspective API** on interactions data ✔

LLMs zero-shot classification

Random, balanced sample of **1,000 messages**

Open-source LLMs deployed locally

Minimum temperature to ensure reproducibility

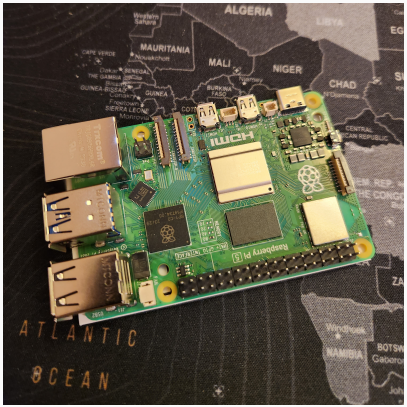
Intel Core i9-14900K CPU

NVIDIA GeForce RTX 4070 Super Windforce OC 12GB GPU

64 GB of RAM (DDR5-4800)

Windows Subsystem for Linux v2.1.5.0 and Ollama v0.1.44 and v0.3.10

Perspective API Classification on Raspberry Pi 5



We ran Perspective API. The classification implied **3,022 hours** of computing **(entire interactions dataset, not only the sample)**.

Part of the process (41% of the classification time) was deployed on a **Raspberry Pi 5**, a small computer board with an ARM CPU and meagre power consumption. This implied a reduction of **39% in our carbon footprint**, saving almost 80 kg of CO_2.

⋆ The reduction is considering the emissions we would have generated running the entire task on an average cloud computing service.

⋆⋆ Picture of the Raspberry Pi 5 used for classification.

Gold Standard

The sample (n = 1,000) was manually annotated by a team of five high-skilled human coders on Labelbox.

Each message was hand-coded by two different coders of the team, therefore, we have **two ground-truth labels for each sample observation**.

GitHub repository.

README | Code of conduct | CC-BY-4.0 license | GPL-3.0 license

gold-standard-toxicity

Gold Standard for Toxicity and Incivility Project
Annotated Data in Spanish for Toxicity and Insults in Digital Social Networks

coverage 100% | version v0.5.3 | repo status Active | DOI 10.5281/zenodo.12574288 | license GNU GPLv3 | license CC BY 4.0
made with Raspberry Pi 5 | supported by tDL

Overview

This repository contains data sets and materials for a gold standard elaboration on toxicity and incivility in the digital sphere based on human coding to benchmark algorithmic classification tasks with transformers and LLMs. **The labelling progress is indicated in the coverage badge above.**

We are labelling two samples of novel datasets of political digital interactions on Twitter (rebranded as X). The first set comprises almost 5 million data points from three Latin American protest events: (a) protests against the coronavirus and judicial reform measures in Argentina during August 2020; (b) protests against education budget cuts in Brazil in May 2019; and (c) the social outburst in Chile stemming from protests against the underground fare hike in October 2019. We are focusing on interactions in Spanish to elaborate a gold standard for digital interactions in this language, therefore, we prioritise Argentinian and Chilean data. The second set contains more than 31 million messages and more than 9 million interactions between 2010 and 2022, covering the election of members of the first Constitutional Convention in Chile, the drafting process and the referendum in which the proposal was rejected.

This project is generously funded by the **OpenAI Academic Programme**, **2024 FAE-UDP Research Grant**, and partially by the **St Hilda's College Muriel Wise Fund at the University of Oxford**. The Training Data Lab research group also logistically supports this project.

Gold Standard

We used a binary codification performed using the core definitions of Perspective for toxicity as follows:

Human annotation

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation.

☑ Krippendorff's α (bootstrap 1,000x) 0.652 (95% CI $[0.601, 0.698]$). These values are tentatively acceptable (Krippendorff, 2019) and, from our viewpoint, denote the complexity of the case addressed.

🕘 The process involved 5 hours of annotation and 1 hour of revision.

GPTs and Open-Source LLMs for Text-Annotation

OpenAI's GPTs

GPT-4o and GPT-4o mini

⋆ o1-preview o1-mini, released recently on September 12, 2024, are not fully available to all API users.

SOTA open-source LLMs

Llama 3.2, Llama 3.1 and Gemma 2 27B

Slightly outdated open-source LLM

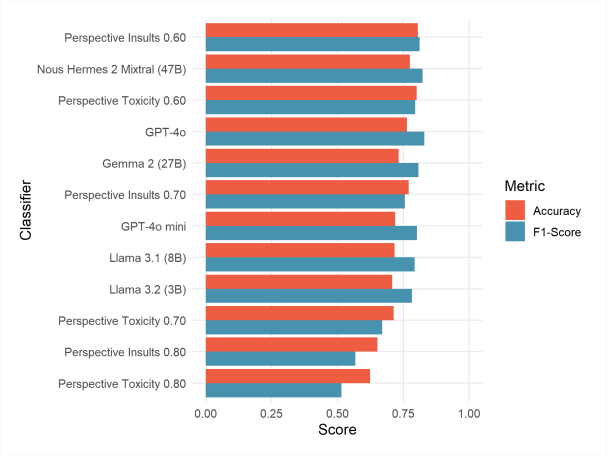
Nous Hermes 2 Mixtral

⋆ Our prompt strategy was the very **same used on the human annotation**.

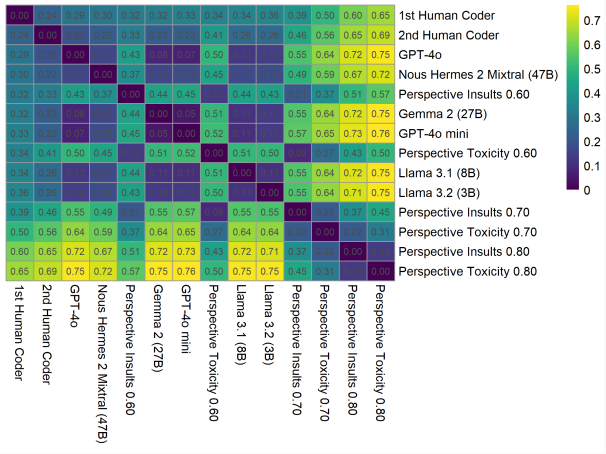
Error Rate Analysis

Perspective at 0.60 (relaxed threshold), **Nous Hermes 2 Mixtral** 47B, and **GPT-4o** show the highest performance in classifying toxicity considering our gold standard.

Llama 3.2, the novel Meta's SOTA, shows one of the lowest performances (still acceptable). However, this is the lightest version since 70 and 405B were beyond our computing infrastructure.



Jaccard Distance between Gold Standard and Classifiers



The heatmap confirms the medium inter-coder reliability of our human annotators. Human coders, **Perspective API at 0.6**, and the best classifiers (i.e., Nous Hermes 2 Mixtral and GPT-4o) tended to be similar.

The same occurs for Perspective indicators at higher thresholds.

Variables and NBRMs

We used two counting indicators: **toxicity** and **insults**. The three main independent variables of interest were **gender**, **indigenous group affiliation**, and **ideological extremism**.

› For ideological extremism, we employ ideological scores estimated using scaling methods (Fábrega, 2022). Scores closer to zero indicate moderate ideological positions, whereas those closer to one reflect greater ideological extremism.

› We fitted naïve NBRMs that incorporated only representatives' age as an additional control. Then, we incorporated squared age to control for nonlinearity and whether the representative was from the Metropolitan Region as a fixed effect.

› We also ran some moderation analyses incorporating interactions and additional models with the **Perspective threshold of 0.7 as robustness checks** ✔.

Results

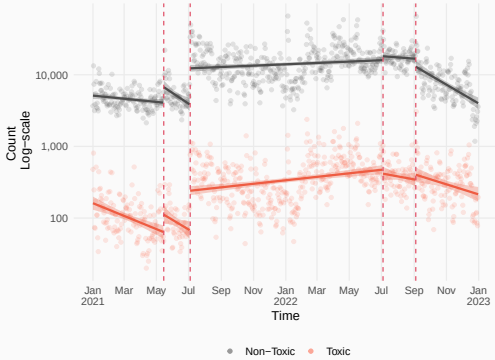
Constitutional Convention

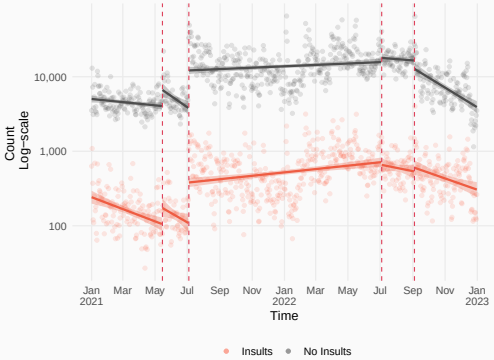
This Convention was held in the framework of the first constitutional process in Chile after massive and violent protests derived from the subway fare hike in October 2019 (Somma et al., 2020; see also Angles et al., 2024).

A new Constitution was understood as a solution to this profound political crisis (Silva, 2023).

⋆ Picture under ⓒ license.

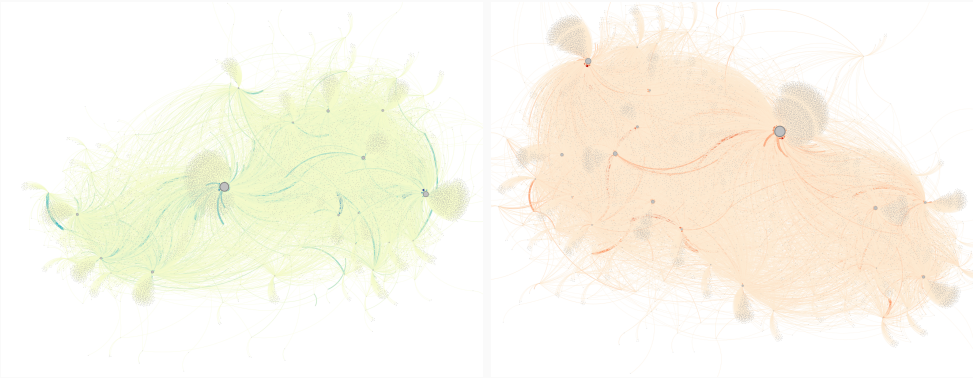
Trends during the Constitutional Convention





⋆ **251,583 toxic messages** and **381,067 insults** during the whole period

Networks during the Constitutional Convention



Toxicity: 62,014 nodes, 153,231 edges, avg. degree 2.471, 16 clusters/communities

Insults: 78,309 nodes, 215,801 edges, avg. degree 2.756, 11 clusters/communities

Determinants of Toxicity and Insults

	Toxic Messages		Insults	
	Model I	Model II	Model III	Model IV
Female	−0.461	−0.612**	−0.412	−0.557*
	(0.313)	(0.297)	(0.311)	(0.295)
Indigenous	0.161	1.029**	0.123	0.983*
	(0.540)	(0.521)	(0.535)	(0.517)
Ideology	1.029*	0.694	0.953*	0.621
	(0.574)	(0.543)	(0.570)	(0.540)
Constant	4.798***	6.216***	5.291***	6.675***
	(0.644)	(0.312)	(0.639)	(0.310)
Age	Yes	Yes	Yes	Yes
Age squared	No	Yes	No	Yes
Region FE	No	Yes	No	Yes
Distribution	NBRM	NBRM	NBRM	NBRM
Theta (θ)	0.304***	0.341***	0.309***	0.346***
VIF	1.025	1.141	1.024	1.145
N	147	147	147	147
Log Likelihood	−1,066.196	−1,054.325	−1,132.966	−1,121.231
AIC	2,142.391	2,122.650	2,275.931	2,256.462

* $p \leq 0.1$; ** $p \leq 0.05$; *** $p \leq 0.01$

Count model coefficients should not be interpreted directly. Therefore, we exponentiated them to obtain rate ratios representing the multiplicative **effect on the expected counts**.

✖ Female representatives received fewer **toxic messages** than their male counterparts

✔ Indigenous representatives indeed receive more **toxic messages**.

Moderation Analyses

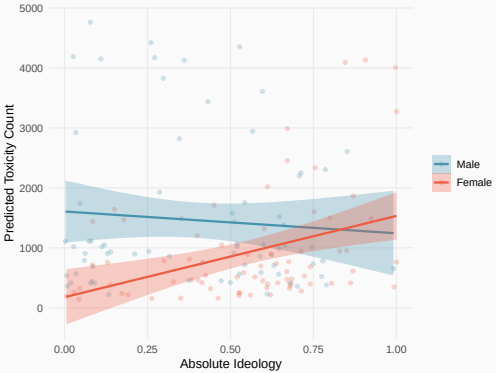
Ideological extremism seemed to act as a **moderator of the effect** of indigeneity on the level of toxicity.

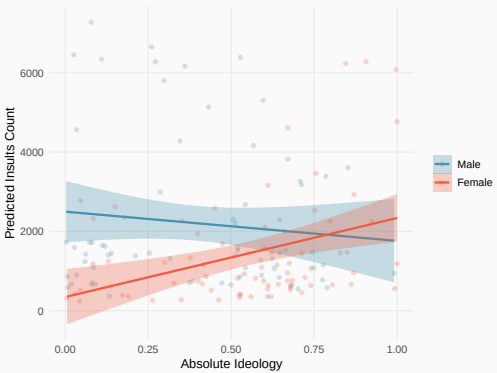
These results hint at nuanced mechanisms by which ideology affects how other key covariates relate to incivility dynamics that require further investigation.

	Toxic Messages		Insults	
	Model I	Model II	Model III	Model IV
Female	−1.257**	−0.726**	−1.211**	−0.671**
	(0.585)	(0.293)	(0.580)	(0.291)
Indigenous	1.081**	3.944**	1.033**	3.871**
	(0.519)	(1.600)	(0.516)	(1.589)
Ideology	−0.113	1.046*	−0.203	0.965*
	(0.750)	(0.550)	(0.744)	(0.547)
Female × Ideology	1.412		1.434	
	(1.061)		(1.053)	
Indigenous × Ideology		−5.354**		−5.277**
		(2.330)		(2.314)
Constant	6.519***	6.089***	6.984***	6.552***
	(0.373)	(0.311)	(0.370)	(0.309)
Age	Yes	Yes	Yes	Yes
Age squared	No	Yes	No	Yes
Region FE	No	Yes	No	Yes
Distribution	NBRM	NBRM	NBRM	NBRM
Theta (θ)	0.343***	0.351***	0.348***	0.356***
VIF	1.152	1.187	1.156	1.190
N	147	147	147	147
Log Likelihood	−1,053.558	−1,051.096	−1,120.424	−1,118.072
AIC	2,123.115	2,118.192	2,256.849	2,252.144

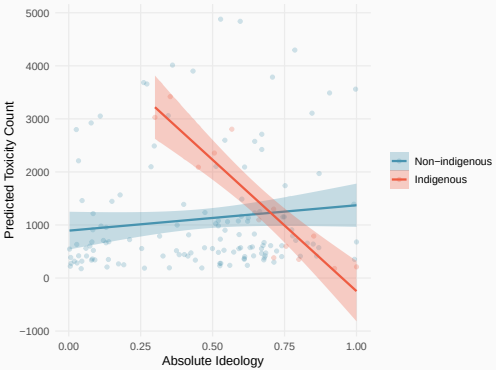
* $p \leq 0.1$; ** $p \leq 0.05$; *** $p \leq 0.01$

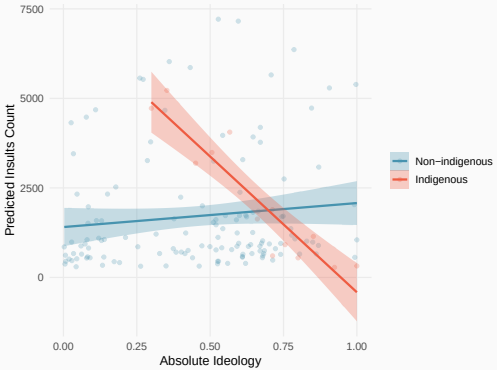
Predicted Toxicity and Insults by Ideology and Gender





Predicted Toxicity and Insults by Ideology and Indigenous Representatives





Takeaways

Takeaways

👍 We **partially accept** the **Indigenous Hypothesis** because these representatives indeed received a higher count of toxic messages but not insults.

👎 We **reject** the **Gender Hypothesis**. Females tended to receive fewer toxic messages.

👎 We **reject** the **Ideology Hypothesis**. Ideology is a moderator rather than a variable mainly affecting toxic messages and insults. This opens **new avenues for future research**.

☑ **Perspective API** using a laxer threshold, **Nous Hermes 2 Mixtral**, and **GPT-4o** are the closest classifiers to our gold standard with human annotators.

› A limitation of this work is our **data's observational nature** and the complexity of the toxicity phenomenon.

Acknowledgements and Funding

This work was supported by OpenAI's Academic Programme; the Universidad Diego Portales under the Faculty of Administration and Economics Research Grant 2024; and the Universität Zürich under the Digital Democracy Workshop 2024 Travel Grant. We also thank the Training Data Lab research group for their support, especially Diego Aguilar, Carla Cisternas, Jaquelin Morillo and Sofía Carrerá-Martínez.

References

References I

Angles, R., Bro, N., Donoso-Guzmán, I., Luna, J. P., Hogan, A., Reutter, J., Rosales-Méndez, H., and Toro, S. (2024). Telar and TelarKG: Data-Driven Insights into Chile's Constitutional Process. *Communications of the ACM*, 67(8):78–79. https://doi.org/10.1145/3653312.

Erikson, J., Håkansson, S., and Josefsson, C. (2023). Three Dimensions of Gendered Online Abuse: Analyzing Swedish MPs' Experiences of Social Media. *Perspectives on Politics*, 21(3):896–912. https://doi.org/10.1017/S1537592721002048.

Esposito, E. and Breeze, R. (2022). Gender and politics in a digitalised world: Investigating online hostility against UK female MPs. *Discourse & Society*, 33(3):303–323. https://doi.org/10.1177/09579265221076608.

Fábrega, J. (2022). Ordenamiento Ideológico en la Convención Constitucional Chilena. *Revista de Ciencia Politica*, 42(1):127–151. https://doi.org/10.4067/s0718-090x2022005000106.

Gorrell, G., Bakir, M. E., Greenwood, M. A., Roberts, I., and Bontcheva, K. (2019). Race and Religion in Online Abuse towards UK Politicians: Working Paper. Preprint, arXiv. https://doi.org/10.48550/arXiv.1910.00920.

References II

Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., and Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, 9(18). https://doi.org/10.1140/epjds/s13688-020-00236-9.

Kim, T. (2023). Violent political rhetoric on Twitter. *Political Science Research and Methods*, 11(4):673–695. https://doi.org/10.1017/psrm.2022.12.

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc. https://doi.org/10.4135/9781071878781.

Orchard, X., Saldaña, M., Pavez, I., and Lagos, C. (2023). 'Does she know how to read?' An intersectional perspective to explore Twitter users' portrayal of women Mapuche leaders. *Information, Communication & Society*, 26(13):2554–2574. https://doi.org/10.1080/1369118X.2023.2252895.

Pérez-Arredondo, C. and Graells-Garrido, E. (2021). Twitter and abortion: Online hate against pro-choice female politicians in Chile. *Journal of Language Aggression and Conflict*, 9(1):127–154. https://doi.org/10.1075/jlac.00056.per.

References III

Silva, P. (2023). Boric y la centroizquierda: de la denostración a la alianza táctica. In Peña, C. and Silva, P., editors, *El gobierno de Gabriel Boric: Entre refundación y reforma*. Catalonia, Santiago.

Somma, N. M., Bargsted, M., Disi Pavlic, R., and Medel, R. M. (2020). No water in the oasis: the Chilean Spring of 2019–2020. *Social Movement Studies*, 20(4):495–502. https://doi.org/10.1080/14742837.2020.1727737.

Theocharis, Y., Barberá, P., Fazekas, Z., and Popa, S. A. (2020). The Dynamics of Political Incivility on Twitter. *SAGE Open*, 10(2). https://doi.org/10.1177/2158244020919447.

Weaving, M., Alshaabi, T., Arnold, M. V., Blake, K., Danforth, C. M., Dodds, P. S., Haslam, N., and Fine, C. (2023). Twitter misogyny associated with Hillary Clinton increased throughout the 2016 U.S. election campaign. *Scientific Reports*, 13(1):5266. https://doi.org/10.1038/s41598-023-31620-w.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub